Multi-Task Learning U-Net for Functional Shoulder Sub-Task Segmentation

En-Ping Chu, Kai-Chun Liu, Chia-Yeh Hsieh, Chih-Ya Chang, Yu Tsao, and Chia-Tai Chan

***Abstract*— The assessment of a frozen shoulder (FS) is critical for evaluating outcomes and medical treatment. Analysis of functional shoulder sub-tasks provides more crucial information, but current manual labeling methods are time-consuming and prone to errors. To address this challenge, we propose a deep multi-task learning (MTL) U-Net to provide an automatic and reliable functional shoulder sub-task segmentation (STS) tool for clinical evaluation in FS. The proposed approach contains the main task of STS and the auxiliary task of transition point detection (TPD). For the main STS task, a U-Net architecture including an encoder-decoder with skip connection is presented to perform shoulder sub-task classification for each time point. The auxiliary TPD task uses lightweight convolutional neural networks architecture to detect the boundary between shoulder sub-tasks. A shared structure is implemented between two tasks and their objective functions of them are optimized jointly. The fine-grained transition-related information from the auxiliary TPD task is expected to help the main STS task better detect boundaries between functional shoulder sub-tasks. We conduct the experiments using wearable inertial measurement units to record 815 shoulder task sequences collected from 20 healthy subjects and 43 patients with FS. The experimental results present that the deep MTL U-Net can achieve superior performance compared to using single-task models. It shows the effectiveness of the proposed method for functional shoulder STS. The code has been made publicly available at https://github.com/RobinChu9890/MTL-U-Net-for-Functional-Shoulder-STS.**

***Clinical Relevance*— This work provides an automatic and reliable functional shoulder sub-task segmentation tool for clinical evaluation in frozen shoulder.**

I. Introduction

Frozen shoulder (FS) is a shoulder condition associating with pain and stiffness that often limits the function of shoulder and the ability for daily living [1, 2]. Assessment of FS is critical for evaluating outcomes to clinical intervention, treatment, and follow up progress. In recent years, several works [3-5] employed wearable sensors to analyze kinematic movements in patients with FS while they perform daily shoulder tasks. These studies demonstrated that wearable-based measurement approach could catch the movement characteristics of affected shoulders and provide objective and quantitative scales for clinicians. Particularly, Lu et al. [6] have suggested that the complex functional shoulder tasks could be divided into a series of sub-tasks for further assessment. The extracted characteristics and kinematics from sub-tasks provide kinematic information related to impaired shoulders in patients with FS compared to that from the complete shoulder task. However, the current annotation of shoulder sub-tasks from the continuous streaming data still relies on manual observation and operation, which is exhausting and may serve as the main barrier in clinical practice. Moreover, the manual errors and inter-observer variability affects the reliability of the obtained kinematics from the sub-tasks [6, 7]. Therefore, an automatic and reliable functional shoulder sub-task segmentation (STS) model is needed to support clinical evaluation and to alleviate burden of clinical professionals.

Conventional human movement segmentation approaches mainly utilized sliding window techniques to divide a successive sensing data into several windows and identify these windows with machine learning classifiers [8-10]. The prior study [7] has shown its feasibility in functional shoulder STS in healthy subjects and patients with FS, which achieves acceptable 83.23% F1-score. Nevertheless, the selection of suitable window size and overlapping percentage is still a challenging and time-consuming process during the development. Furthermore, the larger overlapping ratio often results in high computational complexity [11]. In order to tackle the aforementioned challenges, the advanced semantic segmentation models based on deep learning (DL) can serve as viable approaches to support efficiently automatic movement segmentation. DL-based semantic segmentation models have been successfully developed in the applications of medical image [12, 13] and audio signal processing [14]. This approach can classify each image pixel or data frame as an object category with DL methods [15, 16]. The critical advantage is that it allows to directly process the whole raw data without any framing, and effectively catch global information [17].

One of the popular DL-based semantic segmentation is fully convolutional network (FCN) [12]. An FCN only contains convolutional layers but no fully connected layer, which enables it to process images with arbitrary size and produce a segmentation map with the same size at the final layer. Among variants of FCN, U-Net [13] is one of the most well-known networks for biomedical image segmentation. The architecture of U-Net comprises two parts, a contracting path as an encoder and an expansive path as a decoder. The contracting path follows a typical layout of convolutional neural network (CNN) without fully connected layers to extract the spatial context of input images. The expansive path has a symmetric structure of contracting path, but max-pooling

E.-P. Chu, and C.-T. Chan are with the Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei 112304, Taiwan.

C.-Y. Chang is with the Department of Physical Medicine and Rehabilitation, School of Medicine, National Defense Medical Center, Tri-Service General Hospital, Taipei 114202, Taiwan; Department of Physical Therapy and Assistive Technology, National Yang Ming Chiao Tung University, Taipei 112304, Taiwan.

C.-Y. Hsieh is with the Bachelor's Program in Medical Informatics and Innovative Applications, Fu Jen Catholic University, New Taipei City 242062, Taiwan.

K.-C. Liu and Y. Tsao are with the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei 115201, Taiwan.

(Corresponding author: Y. Tsao, e-mail: yu.tsao@citi.sinica.edu.tw; E.-P. Chu, robinchu.be11@nycu.edu.tw).

layers are replaced with up-convolutional layers to increase feature map dimensions. Several works have applied U-Net to activity recognition and human movement segmentation [18, 19]. Nonetheless, the limited training data is still a challenge restricting the effectiveness of U-Net in wearable-based activity recognition and segmentation, due to the numerous parameters of U-Net.

Multi-task learning (MTL) has been widely applied to improve performance, generalization and accuracy by combining related/different tasks for jointly learning. Especially, MTL has the potential to improve the performance of neural networks with limited training data [20]. Several researches have successfully employed MTL in activity recognition and segmentation [21-23]. For example, Peng et al. [21] developed MLT-based approaches to jointly recognize simple and complex activities, which mined the commonalities and differences between them to improve recognition accuracy. Chen et al. [22] presented a deep multi-task learning based activity and user recognition (METIER) model to solve activity recognition and user recognition jointly, which used soft sharing architecture and attention mechanism to exploit important features simultaneously and transfer knowledge across tasks.

Inspired by the advantages of DL-based semantic segmentation and MTL, we propose a deep MTL U-Net for functional shoulder STS to support FS assessment, which involves the main task of STS and auxiliary task of transition point detection (TPD). For the main STS task, a U-Net architecture including encoder-decoder with skip connection is presented to perform shoulder sub-task classification for each time point. For auxiliary TPD task, a lightweight convolutional neural networks (CNN) architecture is employed to detect the boundary between shoulder sub-tasks, which shares the parameters with the STS encoder. The knowledge extracted from TPD is expected to preserve critical boundary information that could help main task execute more precise transition detection and achieve better STS.

The main contribution is summarized as follows: (i) To provide an automatic and reliable functional shoulder STS tool for clinical evaluation in FS, we propose a deep MTL U-Net model to boost STS performance with the auxiliary TPD task. (ii) We evaluate the proposed model on healthy subjects and patients with FS and compare it with the STS performance using single-task models. (iii) The experimental results demonstrate the effectiveness of the proposed method for functional shoulder STS.

II. Proposed Method

A. Data Preprocessing

Before fed into the network, all time-serial data is first denoised with simple moving average (SMA) filter by averaging a group of samples [24]. A 5-point SMA filter is calculated as (1), where $\tilde{x}$ is the filtered sample, x is the original sample, and n is the sample index. The example smoothed time-serial data is presented in Fig. 1 (a).

$$\tilde{x}_n = \frac{x_{n-2}+x_{n-1}+x_n+x_{n+1}+x_{n+2}}{5} \tag{1}$$

Next, we apply zero-padding to each filtered time-serial sequence $\tilde{X}$ for length resizing [25]. This resizing process

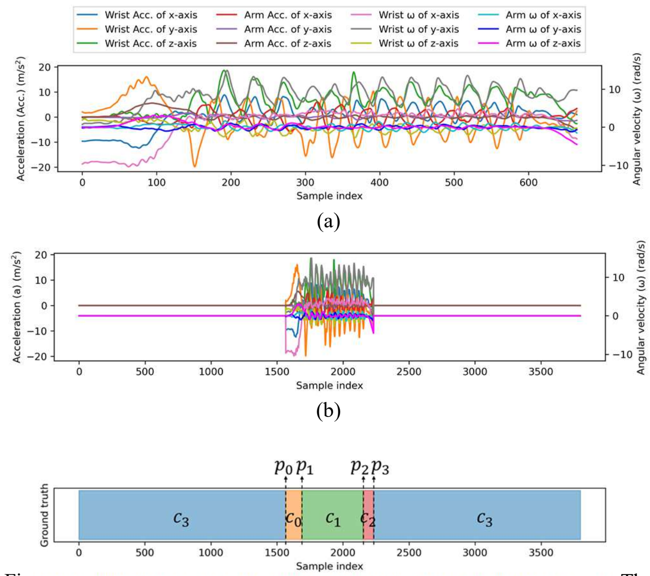

Figure 1. Illustration of time-serial data with data preprocessing. (a) The filtered time-serial data $\tilde{X}$. (b) The time-serial data with zero padding $\hat{X}$. (c) The corresponding class label sequence C and transition points set P for $\hat{X}$.

ensures all sequences to the same size. Let l_i be the length of sequence i, and l_{max} is the maximum of $\{l_i|\forall i \in [1,n]\}$, where n is the total number of sequences. Zero values are added before and after each original time-serial sequence to ensure the new sequence $\hat{X}$ have the same length equal to l_{max}, as shown in Fig. 1 (b).

The added zero samples are labeled as a new sub-task class to be distinguished from the original shoulder sub-task samples. The sub-task boundaries in each IMU sequence are normalized with a respect to l_{max}. The class label sequence $C = \{c_i|\forall c_i \in L, i \in [1, l_{max}]\}$ and the transition points set $P = \{p_j|\forall p_j \in [1, l_{max}],\ j \in [1, n_p]\}$ of resized sequence $\hat{X}$ are illustrated in Fig. 1 (c), where L is the sub-task class set, and n_p is the number of sub-task boundary

B. Deep MTL U-Net

Fig. 2 presents the architecture of the proposed deep MTL U-Net. The structure can be separated into three parts: the STS encoder G_e, the STS decoder G_d, and the transition point detector G_t. G_e and G_d perform sub-task classification on each time point for the STS task while G_e and G_t perform the TPD task. Both tasks share the parameters of G_e.

G_e contains recurrent union of two convolutional layers with kernel size of 1-by-3 and one max-pooling layer with kernel size as 1-by-2. The number of convolutional kernels is doubled after each max-pooling layer. Padding as 1 and stride as 1 are set for convolutional layers to maintain the sequence length. The final contextual encoding is next passed to G_d and G_t respectively.

G_d has a symmetry structure of the encoder, but max-pooling layers are replaced with up-convolutional layer having kernel size as 1-by-2 to increase length and cut channel number in half. After up-convolution, the feature is concatenated with the sequence from corresponding encoder layer to conserve extracted spatial characteristics. For the last layer, a convolutional layer with kernel size as 1-by-1 is used for mapping the feature sequences to the class number.

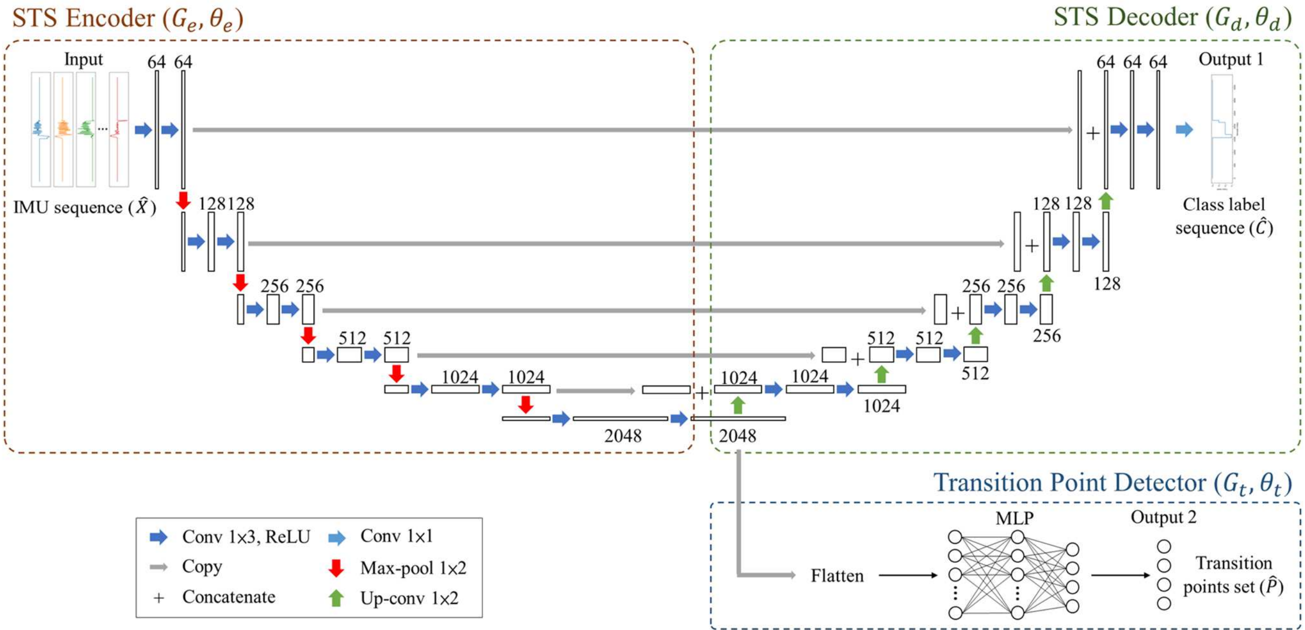

Figure 2. The architecture of proposed deep MTL U-Net

G_t uses multilayer perceptron (MLP) as the main component. We flatten the contextual encoding from the output of G_e, and then input them to two fully connected layers with 210 and 4 neurons.

Given an IMU sequence with zero padding $\hat{X} = \{\hat{x}_i^{m,a} | \forall i \in [1, l_{max}], m \in M, a \in A\}$, where $\hat{x}_i$ is the sample point at time point i, M is the modality set, and A is the axis set. The corresponding class label sequence C and the transition points set P are determined as the target for STS and TPD respectively. The predicted class label sequence $\hat{C} = \{\hat{c}_i | \forall \hat{c}_i \in L, i \in [1, l_{max}]\}$ from G_d and the predicted set of transition points $\hat{P} = \{\hat{p}_j | \forall \hat{p}_j \in [1, l_{max}],\ j \in [1, n_p]\}$ from G_t are given formally by (2) and (3):

$$\hat{C} = G_d(G_e(\hat{X};\ \theta_e); \theta_d), \quad (2)$$

$$\hat{P} = G_t(G_e(\hat{X};\ \theta_e); \theta_t), \quad (3)$$

where θ_e, θ_d, and θ_t are the parameters of G_e, G_d, and G_t, respectively.

The total loss $Loss_{total}$ contains two losses $Loss_{STS}$ and LF_{TPD}, which are given formally by (4)-(6):

$$Loss_{total} = \lambda Loss_{STS} + (1 - \lambda) Loss_{TPD}, \quad (4)$$

$$Loss_{STS} = CE(\hat{C}, C), \quad (5)$$

$$Loss_{TPD} = MSE(\hat{P}, P), \quad (6)$$

where λ is the weight to balance $Loss_{STS}$ and $Loss_{TPD}$, CE is cross entropy loss function, and MSE is mean square error loss function.

III. Experiments

A. Data Collection

This study collects a dataset containing time sequences of functional shoulder tasks collected from IMUs. We recruit 63 subjects, including 20 healthy subjects (10 males, 17 right-handedness, age: 24.55±3.76 years old, height: 168.60±6.73 cm, weight: 67.95±15.34 kg) and 43 patients with FS (16 males, 18 right side affected, 7 both sides affected, age: 57.63±10.58 years old, height: 171.77±47.91 cm, weight: 63.10±11.38 kg). Five functional shoulder tasks are selected from the Shoulder Pain and Disability Index (SPADI) questionnaire [26], containing washing head (WH), washing upper back (WUB), washing lower back (WLB), putting an object on a high shelf (POH), and removing an object from the back pocket (ROB). The data collection is approved by the institutional review board (TSGHIRB No.: A202005024) at the university hospital. All subjects are provided informed consent and voluntary for participation.

Each task is performed once in one recording session and is divided into three shoulder sub-tasks. A total 815 shoulder task sequences are recorded in this study, where 100 sequences performed by healthy subjects and 715 sequences performed by patients at their first and follow-up visits. The longest sequence length l_{max} is 3798. The shoulder sub-task description of five selected shoulder tasks is shown in Table I. Sub-task 1, 2, and 3 of different tasks are trained as the same class to validate the generality of the proposed method. Two IMUs (APDM Inc., Portland, USA) with sampling rate of 128 Hz are fastened to the wrist and upper arm of the dominant side for healthy subjects and the affected side for patients. Each IMU contains a tri-axial accelerometer (range: ±16 g, resolution: 14 bits) and a tri-axial gyroscope (range: ±2000 °/s, resolution: 16 bits) to collect time-serial data with 4 modalities, including acceleration of wrist, angular velocity of wrist, acceleration of upper arm, and angular velocity of upper arm, and 3 axes for $\hat{x}_i$.

B. Implementation Details

The element number of sub-task class set L is four, including three functional shoulder sub-task and one zero-padding class. The number of sub-task boundary n_p is four. The optimizer is AdamW with an initial learning rate of 0.001.

TABLE I. SHOULDER SUB-TASK DESCRIPTION OF FIVE SHOULDER TASK

Task	Sub-task	Description
WH	1	Lift up both hands toward the head
	2	Wash head for a few seconds
	3	Put down both hands and return to the initial position
WUB	1	Lift up the dominant / affected hand toward the upper of back
	2	Wash upper back for a few seconds
	3	Put down the dominant / affected hand and return to the initial position
WLB	1	Lift up the dominant / affected hand toward the lower of back
	2	Wash lower back for a few seconds
	3	Put down the dominant / affected hand and return to the initial position
POH	1	Lift up the dominant / affected hand toward a high shelf while holding a smartphone
	2	Hold the hand for a few seconds
	3	Put down the dominant / affected hand and return to the initial position
ROB	1	Putting a smartphone from the initial position to the back pocket with the dominant / affected hand
	2	Hold the hand for a few seconds
	3	Removing the smartphone from the back pocket to the initial position with the dominant/affected hand

A total 128 epochs are used for mini-batch training, where the batch size is 64.

This work utilizes 10-fold cross validation on the collected dataset for performance evaluation. Three common metrics are chosen as the criteria for performance evaluation, including recall, precision, and F1-score.

The experiments are processed and examined on python 3.9 in a Windows 11 environment with a GPU of NVIDIA RTX 3080. The deep learning network is programed using PyTorch 1.12.1 with CUDA 11.6. The code is available at https://github.com/RobinChu9890/MTL-U-Net-for-Functional-Shoulder-STS.

C. Experimental Results

The results using different weight value (λ) for the deep MTL U-Net are depicted in Fig. 3. When the weight value λ is 0.1 for LF_{STS} and LF_{TPD}, both STS and TPD tasks can reach their best performance of 89.92% and 87.88% in F1-score, respectively.

To demonstrate the effectiveness of the proposed deep MTL U-Net for STS, we compare the proposed method with baseline models without MTL, including a single-task U-Net for STS and a single-task CNN for TPD. These simplified networks have the same experimental details, and their parameters are optimized. The experiment results are presented in Table 2. It shows that the proposed deep MTL U-Net can reach superior performance to the single-task models. The F1-score on STS and TPD are increased by 0.66% and 0.32%, respectively. Moreover, the segmentation f1-score (89.92%) of the proposed model notably outperforms that

TABLE II. THE PERFORMANCE COMPARISON BETWEEN DEEP MTL U-NET AND SIMPLIFIED MODELS

Task	Structure	Rec[a] (%)	Pre[a] (%)	F1[a] (%)
STS	Deep MTL U-Net	**90.31**	**89.64**	**89.92**
	U-Net only (without G_T)	89.52	89.08	89.26
TPD	Deep MTL U-Net	88.26	87.61	87.88
	CNN only (without G_D)	87.62	87.55	87.56

a. Rec: Recall, Pre: Precision, F1: F1-score

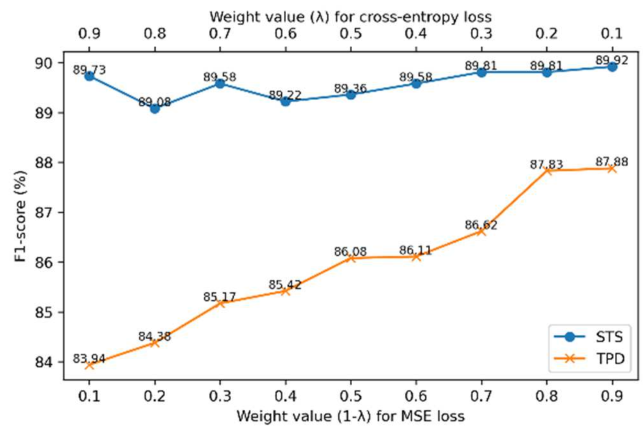

Figure 3. The performance line chart of segmentation decoder and transition points detector in proposed multi-tasks network with different λ.

(83.23%) of the prior study [7] approach using conventional sliding window and machine learning techniques.

When performing STS and TPD tasks jointly, MTL could transfer knowledge across them, which are hard to learn by single-task models. On one hand, for STS task, the fine-grained information of transition from TPD task is considered while performing time-serial segmentation by giving TPD task large weight. It helps STS task have better ability to tackle the boundary between sub-tasks. On the other hand, the TPD task obtains critical contextual knowledge from STS task to efficiently detect transition points between sub-tasks with only a relatively small weight for STS task. Moreover, the differences between STS and TPD tasks are the inductive bias, joint learning and sharing parameters could improve generalization ability of both tasks.

IV. CONCLUSION

In this work, we propose deep MTL U-Net to provide automatic functional shoulder STS for clinical evaluation in FS. We apply an MTL architecture to combining the encoder-decoder U-Net and CNN structures, where the former performs the main STS task and the latter performs the auxiliary TPD task. This study conducts an experiment using wearable IMUs to record 815 functional shoulder tasks, which are collected from 20 healthy subjects and 43 patients with FS. The experimental results present that the proposed methods can achieve superior performance compared to using single-task models. It shows the effectiveness of proposed deep MTL U-Net for functional shoulder STS. Future work aims to exploit other deep learning methods to boost STS performance, such as gated recurrent unit (GRU), and self-attention mechanism. Another intent is to recruit more healthy seniors to reduce the age gap between healthy subjects and FS patients.

REFERENCES

[1] C. Robinson, K. M. Seah, Y. Chee, P. Hindle, and I. Murray, "Frozen shoulder," *The Journal of bone and joint surgery. British volume,* vol. 94, no. 1, pp. 1-9, 2012.

[2] R. Dias, S. Cutts, and S. Massoud, "Frozen shoulder," *Bmj,* vol. 331, no. 7530, pp. 1453-1456, 2005.

[3] S.-H. Lee, S.-C. Yeh, R.-C. Chan, S. Chen, G. Yang, and L.-R. Zheng, "Motor ingredients derived from a wearable sensor-based virtual reality system for frozen shoulder rehabilitation," *BioMed research international,* vol. 2016, 2016.

[4] Y.-C. Lin, Y.-J. Tsai, Y.-L. Hsu, M.-H. Yen, and J.-S. Wang, "Assessment of shoulder range of motion using a wearable inertial sensor network," *IEEE Sensors Journal,* vol. 21, no. 13, pp. 15330-15341, 2021.

[5] H.-C. Lin, S.-Y. Chiang, K. Lee, and Y.-C. Kan, "An activity recognition model using inertial sensor nodes in a wireless sensor network for frozen shoulder rehabilitation exercises," *Sensors,* vol. 15, no. 1, pp. 2181-2204, 2015.

[6] T.-Y. Lu, K.-C. Liu, C.-Y. Hsieh, C.-Y. Chang, Y. Tsao, and C.-T. Chan, "Instrumented shoulder functional assessment using inertial measurement units for frozen shoulder," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021: IEEE, pp. 1-4.

[7] C.-Y. Chang *et al.*, "Automatic functional shoulder task identification and sub-task segmentation using wearable inertial measurement units for frozen shoulder assessment," *Sensors,* vol. 21, no. 1, p. 106, 2020.

[8] Y. Zhang, L. Wang, H. Chen, A. Tian, S. Zhou, and Y. Guo, "IF-ConvTransformer: A Framework for Human Activity Recognition Using IMU Fusion and ConvTransformer," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 6, no. 2, pp. 1-26, 2022.

[9] Y.-W. Kim, W.-H. Cho, K.-S. Kim, and S. Lee, "Inertial-Measurement-Unit-Based Novel Human Activity Recognition Algorithm Using Conformer," *Sensors,* vol. 22, no. 10, p. 3932, 2022.

[10] D. Strömbäck, S. Huang, and V. Radu, "Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 4, no. 4, pp. 1-22, 2020.

[11] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, "A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors," *Sensors,* vol. 19, no. 22, p. 5026, 2019.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015: Springer, pp. 234-241.

[14] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 9, pp. 1570-1584, 2018.

[15] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence,* 2021.

[16] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857,* 2017.

[17] X.-X. Yin, L. Sun, Y. Fu, R. Lu, and Y. Zhang, "U-Net-Based medical image segmentation," *Journal of Healthcare Engineering,* vol. 2022, 2022.

[18] Y. Zhang, Z. Zhang, Y. Zhang, J. Bao, Y. Zhang, and H. Deng, "Human activity recognition based on motion sensor using u-net," *IEEE Access,* vol. 7, pp. 75213-75226, 2019.

[19] N. Yoshimura, T. Maekawa, T. Hara, and A. Wada, "Acceleration-based Activity Recognition of Repetitive Works with Lightweight Ordered-work Segmentation Network," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 6, no. 2, pp. 1-39, 2022.

[20] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796,* 2020.

[21] L. Peng, L. Chen, Z. Ye, and Y. Zhang, "Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 2, no. 2, pp. 1-16, 2018.

[22] L. Chen, Y. Zhang, and L. Peng, "METIER: A deep multi-task learning based activity and user recognition model using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 4, no. 1, pp. 1-18, 2020.

[23] S. Xia, L. Chu, L. Pei, W. Yu, and R. Qiu, "A Boundary Consistency-aware Multi-task Learning Framework for Joint Activity Segmentation and Recognition with Wearable Sensors," *IEEE Transactions on Industrial Informatics,* 2022.

[24] S. W. Smith, "The scientist and engineer's guide to digital signal processing," ed: California Technical Pub. San Diego, 1997.

[25] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," *Journal of Big Data,* vol. 6, no. 1, pp. 1-13, 2019.

[26] J. D. Breckenridge and J. H. McAuley, "Shoulder pain and disability index (SPADI)," *Journal of physiotherapy,* vol. 57, no. 3, pp. 197-197, 2011.